ExtremeRotations In The Wild

Hana Bezalel[1] Dotan Ankri[1] Ruojin Cai[2] Hadar Averbuch-Elor[1,2]

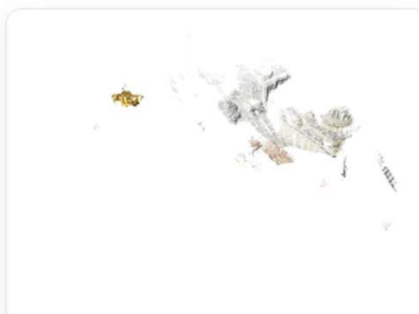
Tel Aviv University[1] Cornell University[2]

CVPR 2025

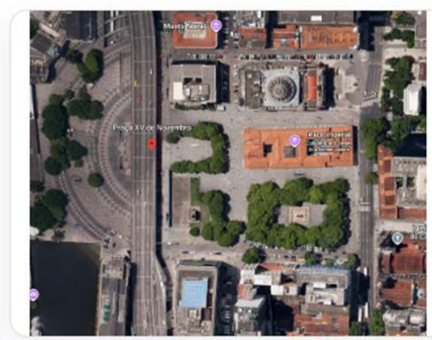
Why Extreme Settings Matter

- 3D reconstructions rely on dense, overlapping views
- What about sparse-view reconstruction?
- Current methods fails when views are sparse (both classical and foundation models)

Input Images

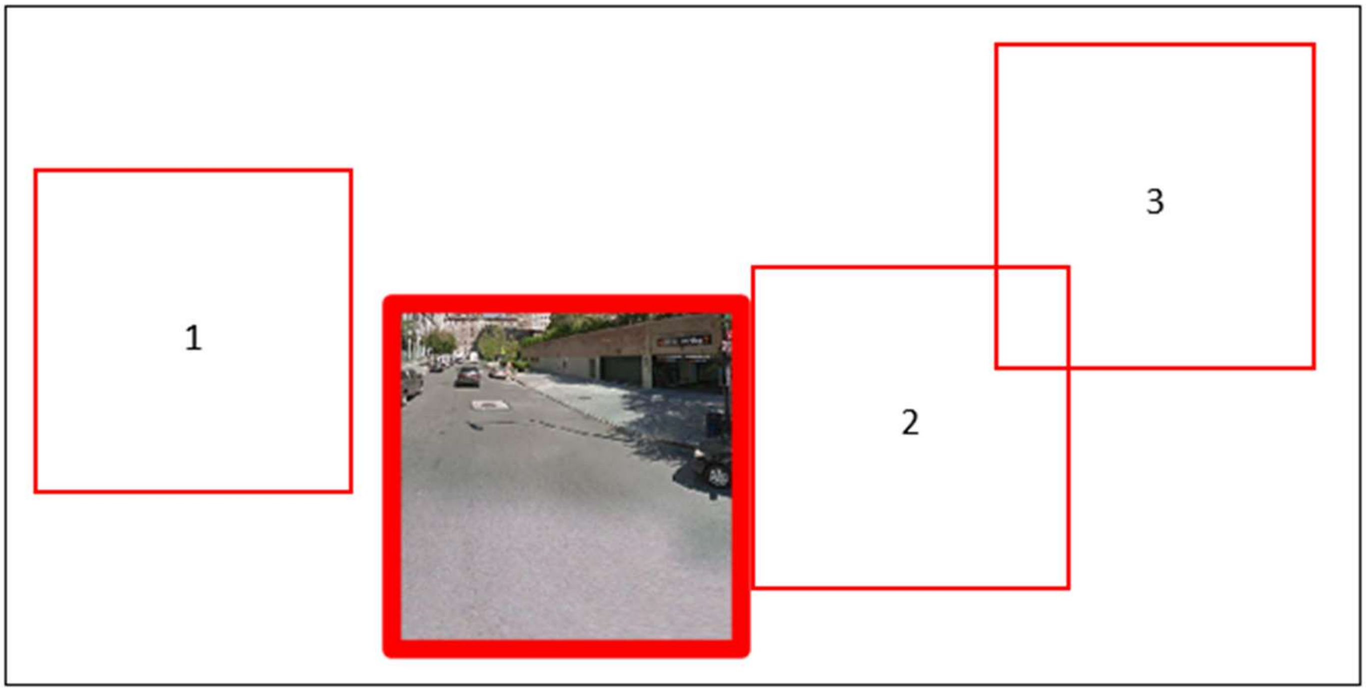


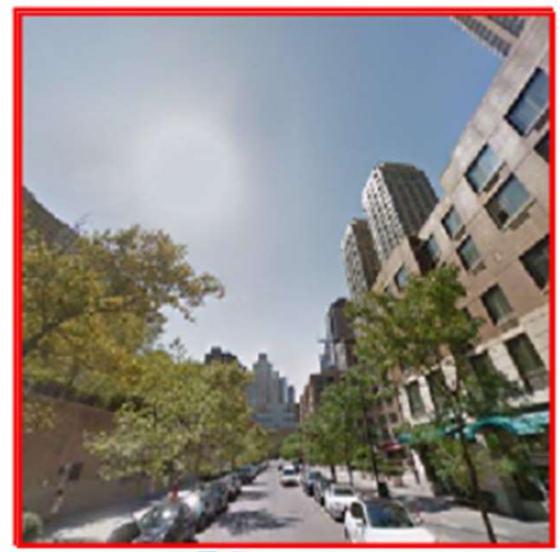
Pretrained π^3



Google Earth

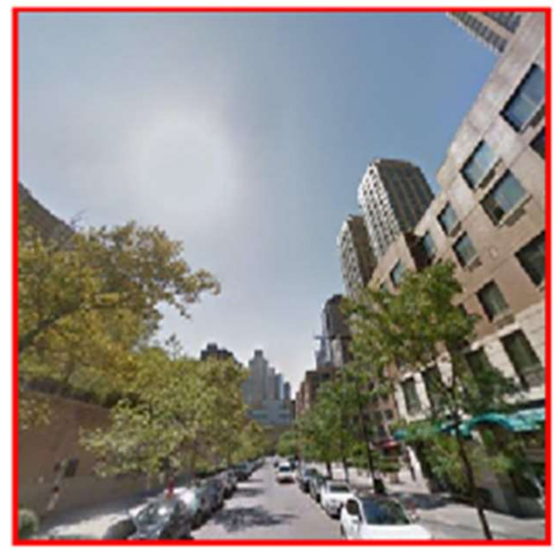
Do we (humans) have a prior that helps predict rotation in extreme non-overlapping image pairs?





?

Do we (humans) have a prior that helps predict rotation in extreme non-overlapping image pairs?



Can we teach a model to exploit these implicit geometric cues?

Prior work

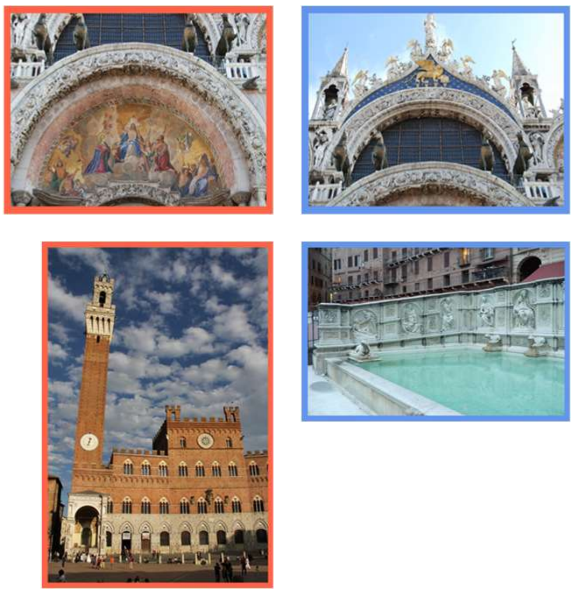
- Emulate perspective images by cropping regions from panoramic views
- Limitations: 90 deg fixed FoV, constrained environments

Ruojin Cai, Bharath Hariharan, Noah Snavely, and Hadar Averbuch-Elor. Extreme rotation estimation using dense correlation volumes. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 2021 ,14575–14566

Shay Dekel, Yosi Keller, and Martin Cadik. Estimating extreme 3d image rotations using cascaded attention. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 7 ,6 ,4 ,3 ,2 ,1 .2024 ,2598–2588

What makes "In-the-Wild" Hard?

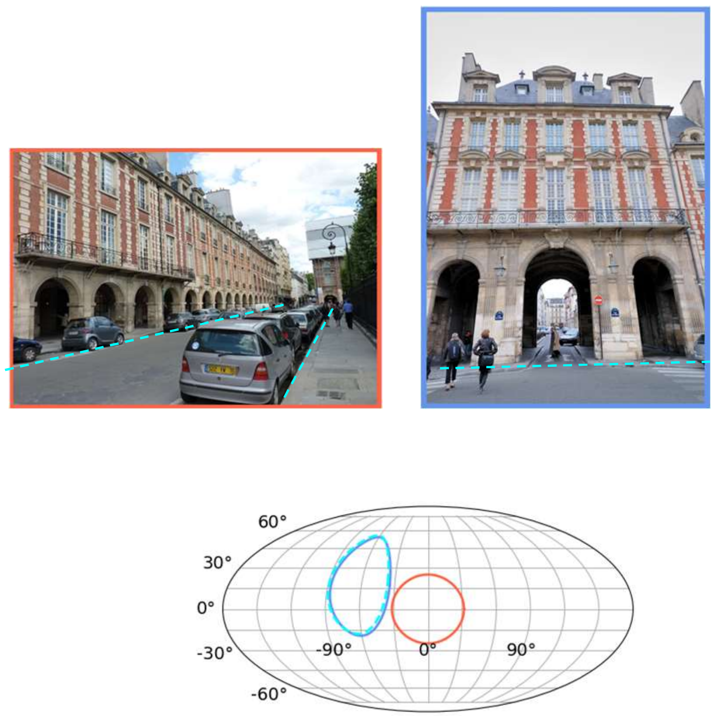
Intrinsic parameters



Occlusions & transient

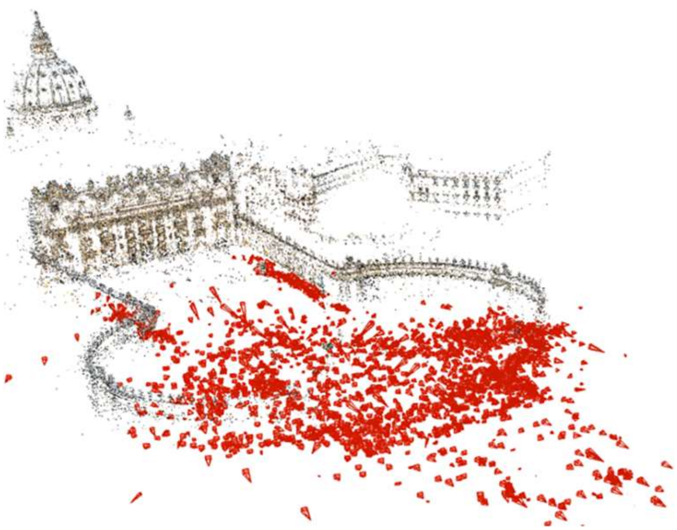
Weather & Time of day

Do we (humans) have a prior that helps predict rotation in extreme non-overlapping in-the-wild image pairs?



ExtremeMegaLandmarks dataset -
Scaling Up Extreme Rotations Pairs

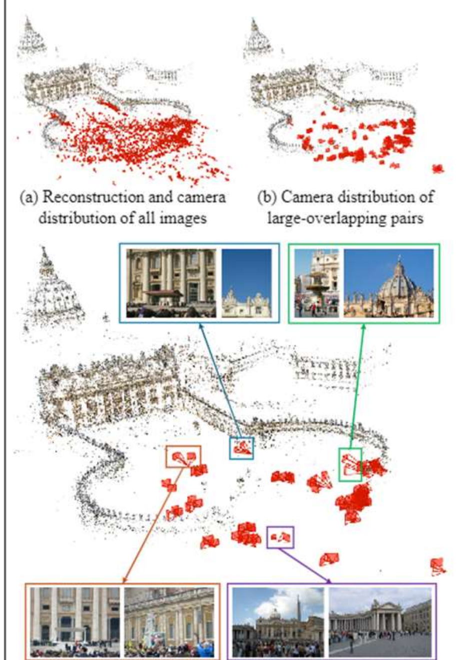
- Internet photo collections - diverse scenes, real-world capture conditions
- Ground truth relative poses obtained via SFM reconstruction
- Identifying pairs with predominant rotational motion
- Pairs span large rotation ranges including opposite facing



Constructing the ExtremeLandmarkPairs Dataset: Challenges in Pair Selection and Balance

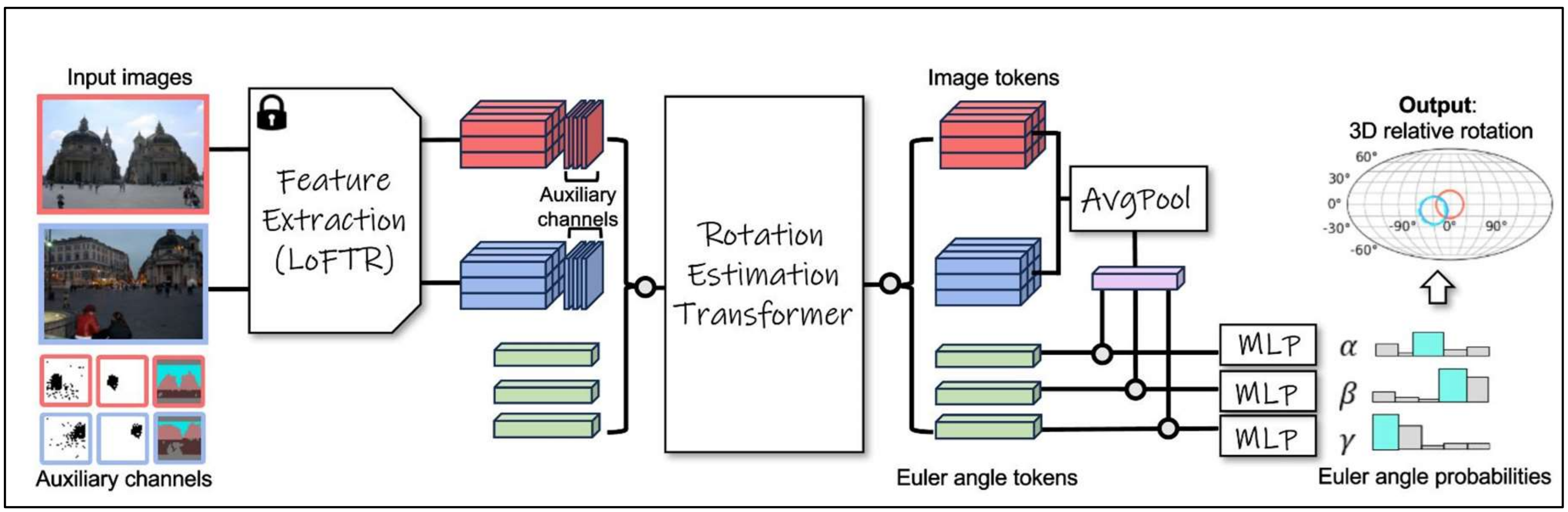
- **Scale ambiguity**:
  SfM poses known only up to per-scene scale
  ➔ Mutual KNN graphs per scene to find spatially close cameras.

- **Photographer bias**:
  Scene-centric collections
  ➔ non-overlapping pairs are inherently rare.


(a) Reconstruction and camera distribution of all images
(b) Camera distribution of large-overlapping pairs
(c) Camera distribution of non-overlapping pairs

Rethinking the Architecture for In-The-Wild Generalization

- LoFTR Encoder
- KP and segmentation masks
- Rotation Estimation Transformer
- 3 learnable tokens



From Controlled to In-The-Wild: Progressive Training Strategy

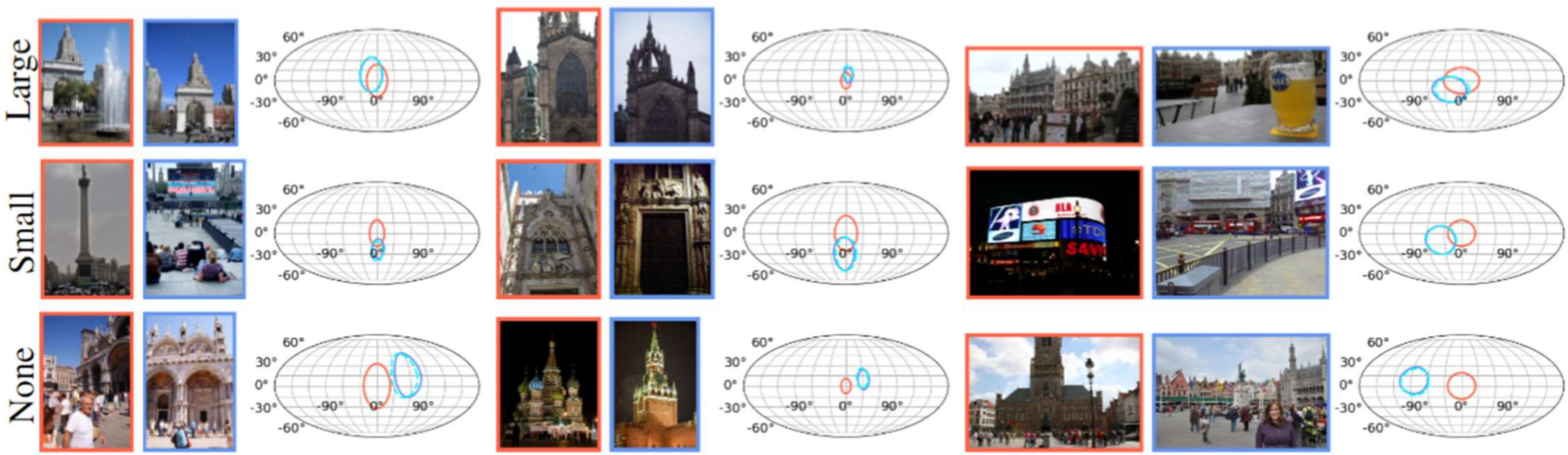
90 deg FoV

FoV augmentations

Image level appearance augmentations

Real Image pairs (ELP dataset)

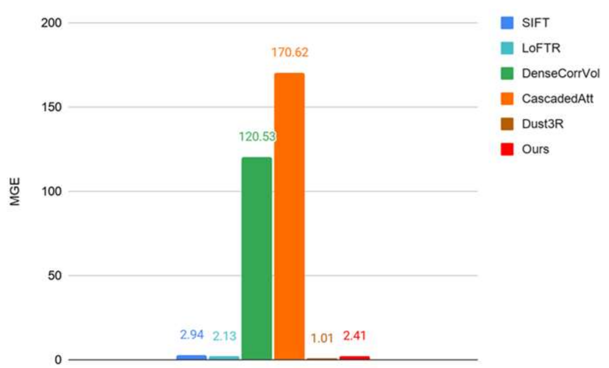
Qualitative Results



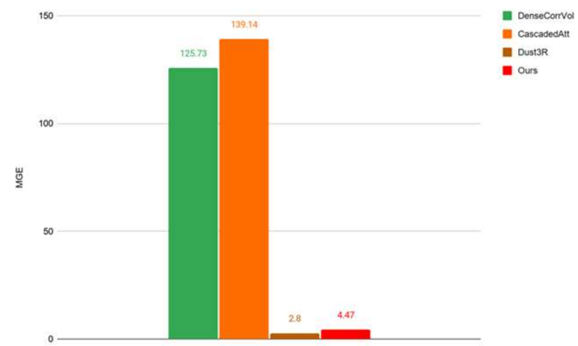
Quantitative Results-MGE

$$\text{err} = \arccos\left(\frac{tr(\mathbf{R}^T\mathbf{R}^*) - 1}{2}\right)$$

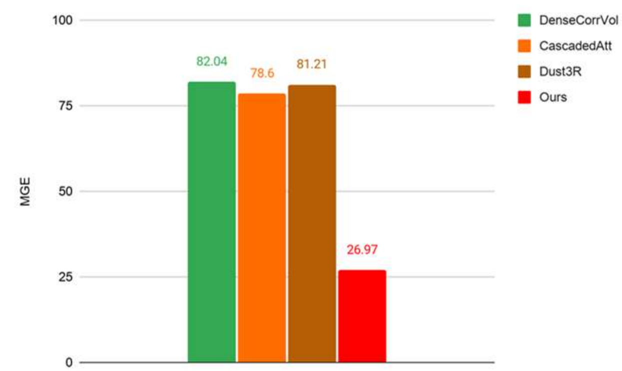
Large



Small



None



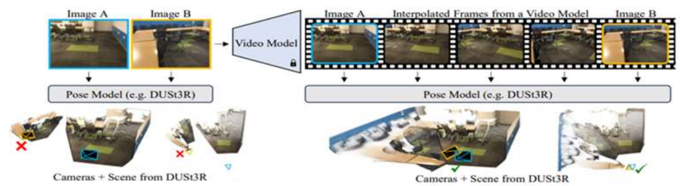
* Dust3R was trained on MegaDepth that wELP was derived from

Follow up work

- **Generative Video for Pose Estimation**

  Use a video generation model to hallucinate intermediate frames between two views,

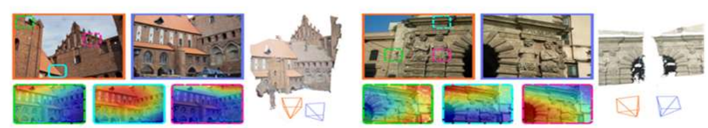
  creating a dense visual transition that simplifies pose estimation.



- **Extreme-View Geometry in 3DFMs**

  3D foundation models already capture extreme-view geometry in their internal representations

  a lightweight bias-only alignment further refines their 3D representations with minimal parameter tuning.



Ruojin Cai, Jason Y Zhang, Philipp Henzler, Zhengqi Li, Noah Snavely, and Ricardo Martin-Brualla. Can generative video models help pose estimation? In Proceedings of the Computer Vision and Pattern Recognition Conference, pages 2 .2025 ,16773–16764

Yiwen Zhang, Joseph Tung, Ruojin Cai, David Fouhey, Hadar Averbuch-Elor. Emergent Extreme-View Geometry in 3D Foundation Models. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 36411-36421 .pp ,2026

Thank you!